

Beyond protein expression, MOPED goes multi-omics

Elizabeth Montague^{1,2,3,4}, Imre Janko^{2,3,4}, Larissa Stanberry^{1,2,3,4}, Elaine Lee^{2,3,4}, John Choiniere^{1,2,4}, Nathaniel Anderson^{1,2,4}, Elizabeth Stewart^{1,4}, William Broomall^{2,3,4}, Roger Higdon^{1,2,3,4}, Natali Kolker^{2,3,4} and Eugene Kolker^{1,2,3,4,5,6,*}

¹Bioinformatics and High-Throughput Analysis Laboratory, Center for Developmental Therapeutics, Seattle Children's Research Institute, Seattle, WA, USA 98101, ²High-Throughput Analysis Core, Seattle Children's Research Institute, Seattle, WA, USA 98101, ³CDO Analytics, Seattle Children's, Seattle, WA, USA 98101, ⁴Data-Enabled Life Sciences Alliance (DELSA Global), Seattle, WA, USA 98101, ⁵Departments of Biomedical Informatics and Medical Education and Pediatrics, University of Washington, Seattle, WA, USA 98109 and ⁶Department of Chemistry and Chemical Biology, College of Science, Northeastern University, Boston, MA 02115

Received September 02, 2014; Revised October 23, 2014; Accepted November 02, 2014

ABSTRACT

MOPED (Multi-Omics Profiling Expression Database; <http://moped.proteinspire.org>) has transitioned from solely a protein expression database to a multi-omics resource for human and model organisms. Through a web-based interface, MOPED presents consistently processed data for gene, protein and pathway expression. To improve data quality, consistency and use, MOPED includes metadata detailing experimental design and analysis methods. The multi-omics data are integrated through direct links between genes and proteins and further connected to pathways and experiments. MOPED now contains over 5 million records, information for approximately 75 000 genes and 50 000 proteins from four organisms (human, mouse, worm, yeast). These records correspond to 670 unique combinations of experiment, condition, localization and tissue. MOPED includes the following new features: pathway expression, Pathway Details pages, experimental metadata checklists, experiment summary statistics and more advanced searching tools. Advanced searching enables querying for genes, proteins, experiments, pathways and keywords of interest. The system is enhanced with visualizations for comparing across different data types. In the future MOPED will expand the number of organisms, increase integration with pathways and provide connections to disease.

INTRODUCTION

When it was created in 2011, MOPED (Multi-Omics Profiling Expression Database) initially focused on just one class of molecules—proteins (1,2). Current high-throughput approaches produce a vast amount of diverse molecular expression data (3) that can be utilized to enable a deeper understanding of biological processes. The need for multi-omics data systems is highlighted by the number of biological discoveries that were made possible by examining multiple omics data types (4–7).

With the addition of transcriptomics in 2014, MOPED became a multi-omics resource (8). Utilizing consistent processing, MOPED provides gene and protein expression data that can be compared across experiments, conditions and tissues. Integration of these multi-omics data occurs through direct links between genes and proteins, connections to pathways and consistent experimental metadata.

MOPED provides detailed experimental metadata via a standardized multi-omics checklist. The checklist provides details about research objectives, experimental design, protocols, instrumentation, data processing and analysis (9–11). This added layer of experimental metadata can be used by researchers to assess data quality, consistency, usability and reproducibility.

New features and data have been added to MOPED to provide meaningful integration of multi-omics. These new features include Pathway Details pages, pathway expression summaries, experimental metadata checklists for all gene and most protein expression experiments, summary statistics for experiments and more advanced searching tools including Boolean searching. MOPED data are integrated in the context of biological pathways through pathway summaries and expression statistics, allowing the user to examine connections and interactions (12). Advanced searching

*To whom correspondence should be addressed. Tel: +206-884-7170; Fax: +206-987-7660; Email: eugene.kolker@seattlechildrens.org

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation, The Robert B. McMillen Foundation or Seattle Children's Research Institute.

enables queries of diverse data types, including genes, proteins, pathways, metadata, conditions, localizations, tissues and organisms.

NEW DATA

MOPED data come from individual researchers and public repositories (Table 1) (13–21), (<http://proteomecentral.proteomexchange.org>, <https://www.proteomicsdb.org/>). Consistent processing of hand-curated experiments provides a foundation for integrated data analysis (1,8,22–23). All proteomics data are processed and analyzed using SPIRE; protein expression data combined with normal tissue protein concentrations are used to uniquely estimate protein concentrations (1,22,24–25). Preprocessed transcriptomics data are downloaded from GEO and analyzed using the R package LIMMA (8,26–27). The addition of experimental metadata provides details about experimental design, instrument, processing and analysis (9,10).

The number of expression records contained in MOPED has increased to over 5 million: ~1 million for protein absolute expression, 100 000 for protein relative expression and over 4 million for gene relative expression data (Table 2). These expression records correspond to ~75 000 genes and 50 000 proteins from four organisms (human, mouse, worm, yeast). MOPED includes data from ~25 000 human genes and 18 000 human proteins. These expression records are from 670 unique combinations of experiment, condition, localization and tissue. Each MOPED experiment can include a number of these unique combinations. The proteins and genes are linked to over 5000 pathways from three prominent pathway resources (BioCyc, Reactome and PANTHER) (16–18).

MOPED is continually expanding by monitoring and downloading new experimental data from public repositories: PRIDE, PeptideAtlas, proteomeXchange, ProteomicsDB and GEO (19–21), (<http://proteomecentral.proteomexchange.org>, <https://www.proteomicsdb.org/>). Recent additions to MOPED include two prominent Human Proteome studies of protein absolute expression data on 30 tissues and cell lines from the Pandey group (28) and on 21 tissues, cell lines and fluids from the Kuster group (Figure 1) (29) (raw data download from ProteomeXchange, PXD000561 and PXD000865). Other recent data additions include absolute expression from deep proteome profiling of 11 cell lines (30) and proteomics analysis of the NCI-60 cell line panel (31). To provide a valuable resource for investigating cancer mechanisms, transcriptomics data from the NCI-60 cell line panel (32) are being incorporated along with the proteomics data.

MOPED FEATURES

The enhanced MOPED interface enables querying of proteins, genes and keywords to access expression data, summary information and advanced data visualization. The interface not only allows exploration of expression data points but also provides pre-calculated pathway statistics, summary pages and visualizations for a comprehensive yet concise view of the data. The home search page features six

tabs: Protein Absolute Expression, Protein Relative Expression, Gene Relative Expression, Pathways, Experiments and Visualization.

Each of the six tabs in MOPED is designed to enable rapid querying. MOPED uses Lucene for full-text indexing, and AspectJ and Google Analytics for tracking and optimization (Apache Foundation, <http://lucene.apache.org/>; Eclipse Foundation, <http://eclipse.org/aspectj/>; Google, <http://www.google.com/analytics/>). The queries are built and executed via the Basic Search (Figure 2A and B) and Advanced Search (Figure 3) options. The Highlights section features preset examples.

Search capabilities

Basic Search now supports wildcards and logic statements. For example, in the Gene Relative Expression tab, a basic search for ‘gene=BRC*’ returns relative expression data for all genes from human and model organisms, whose names start with BRC (Figure 2A). Similarly, a basic search in the Gene Relative Expression tab for ‘gene=BRC* and organism=human and (condition=standard or condition=*cancer*)’ returns all human gene relative expression data for genes’ names that start with BRC and conditions that are either standard or have ‘cancer’ in the name (Figure 2B).

As a multi-omics resource, MOPED supports gene and protein queries under each of the six tabs. For example, a protein ID can be used to search under the Gene Relative Expression tab to obtain all the gene relative expression records for the gene that corresponds to the query protein. Pathways that contain the query protein/gene are conveniently linked to the Results table. This integration enables comprehensive searching with a single ID, whether gene or protein.

The Advanced Search panels available for all tabs have been expanded (Supplementary Figure S1). The drop-down menus show all of the available search options, for example: organism, experiment, condition, localization, tissue, chromosome location, expression ratio and data type. In addition, MOPED allows switching between tabs while preserving the query results.

Results tables

Query results are displayed in tables that are divided into sections for ease of use. For instance, searching under Gene Relative Expression returns a table of gene relative expression records. Each record has an expression ratio and the corresponding *P*-value for each gene in an experiment. This expression ratio reflects a comparison between either tissues or conditions as determined by the experimental design. The expression records are also linked to chromosomes and pathways (14–18). All values can be sorted and downloaded.

Details pages

Detailed data on each biological entity are provided on the Details pages. For example, a Gene Details page contains a link to the corresponding protein, chromosome mappings,

Table 1. External sources for MOPED data

Data type	Expression data	Names, descriptions and mappings
Protein	PRIDE, PeptideAtlas, proteomeXchange, ProteomicsDB, Individual Labs	UniProt
Gene	GEO	Ensembl, Entrez
Pathway		BioCyc, PANTHER, Reactome

Expression data come from public repositories and individual labs (13–21) (<http://proteomecentral.proteomexchange.org>, <https://www.proteomicsdb.org/>).

Table 2. Current data statistics for MOPED as of August 31, 2014

	Protein absolute expression	Protein relative expression	Gene relative expression
Number of expression records	960 574	92 824	4 224 869
Unique combinations of experiment, condition, localization and tissue	362	59	249
Number of experiments	73	18	183
Number of conditions	94	48	272
Number of tissues	137	17	112
Number of human proteins or genes	18 475	7987	24 517
Number of mouse proteins or genes	14 116	3557	26 904
Number of worm proteins or genes	10 178	466	18 143
Number of yeast proteins or genes	5402	0	5811
Number of pathways	5417	3295	5517

Each experiment consists of unique combinations of conditions, localizations and tissues. Expression records are expression values for a protein or gene corresponding to a combination of experiment, condition, localization and tissue.

**Figure 1.** Absolute expression chart for two prominent 'Nature' proteome studies, which highlights the absolute expression of protein P07148 in standard conditions across different tissues (28,29).

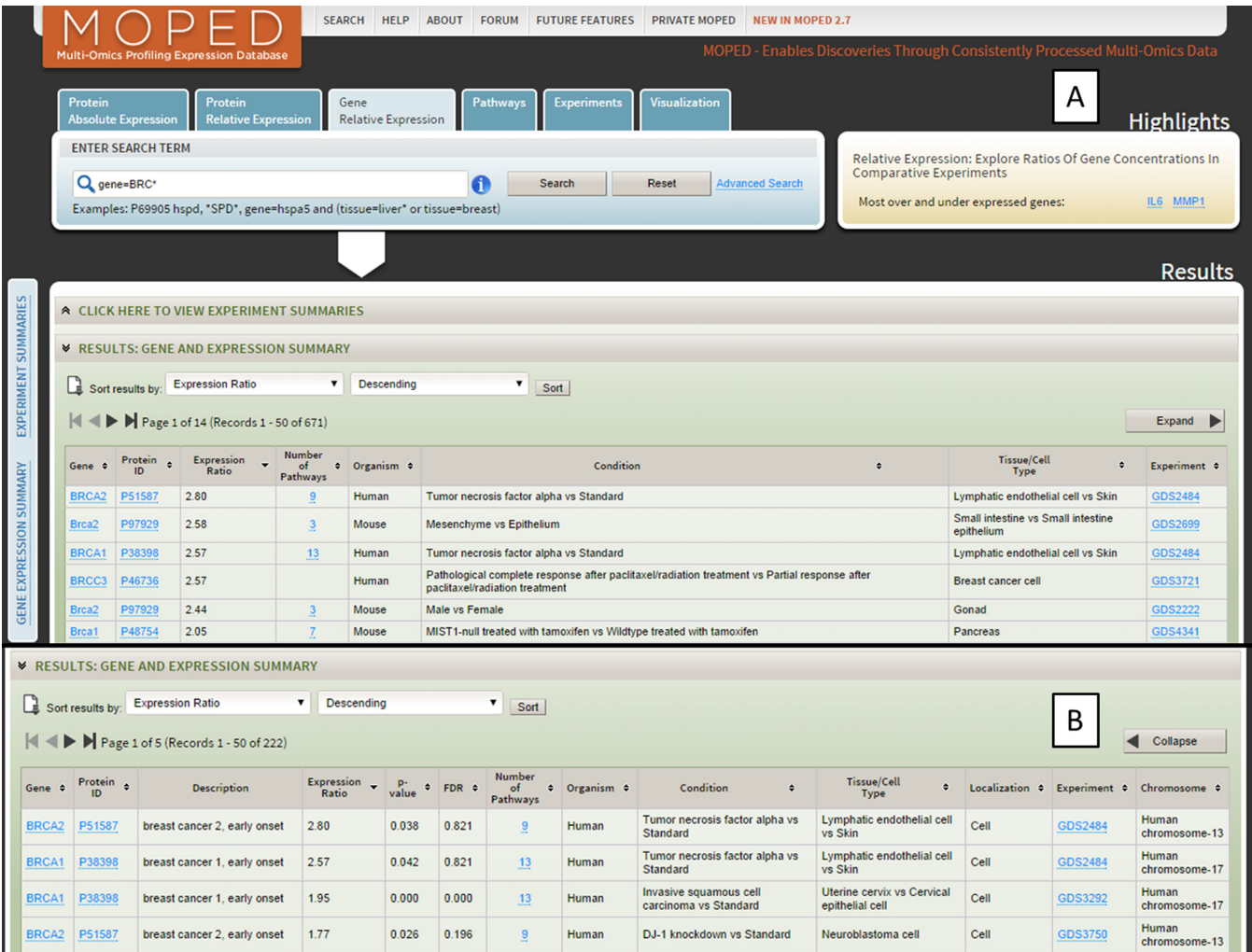


Figure 2. (A) Basic Search example for 'gene = BRC*', default view. (B) Basic Search example for 'gene=BRC*' and organism=human and (condition=standard or condition=*cancer*'), extended view. Search results are expanded to include *P*-value, False Discovery Rate (FDR), localization and chromosome.

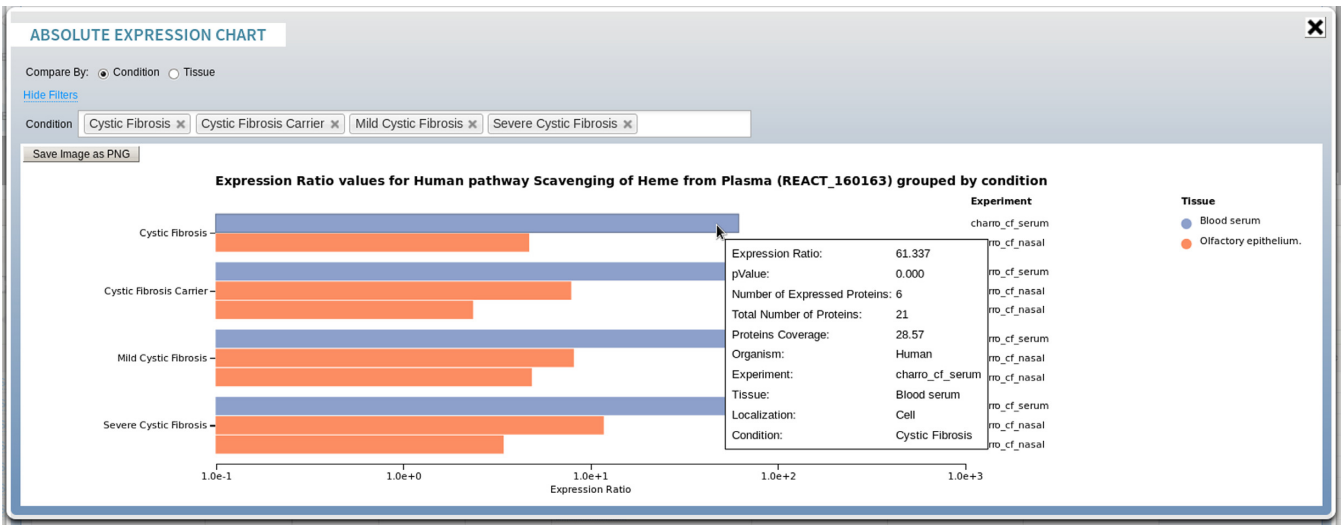


Figure 3. Pathway Absolute Expression bar chart of Scavenging of Heme from Plasma. Chart can be compared and filtered by tissue or condition. Mouse-overs provide more information about the expression data point.

gene description and a link to GeneCards (33). Genes are also linked to external sources including NCBI Gene, GO, etc. and pathways (Reactome, PANTHER and BioCyc) (Table 3) (13–18,33–44). Gene and Protein Details pages display gene and protein expression data, respectively. The pages are enhanced by the embedded expression bar charts.

Details pages for Pathways display pathway absolute expression data and their bar graph together with links to external resources and a list of associated genes and proteins. The associated genes and proteins link directly to their expression data. The Experiment Details page includes the standardized experimental metadata checklist and summary statistics, such as the number of over- or under-expressed proteins or genes (9,10).

Pathways

MOPED adopted a pathway-centric view to integrate gene and protein expression data. The new Pathways tab enables querying for pathway names, keywords and biomolecules in addition to the standard organism, experiment, condition, localization and tissue options. Protein absolute expression data from a given experiment are mapped onto established pathways from Reactome, PANTHER and BioCyc (16–18). Using this mapping, MOPED provides summary statistics for pathway expression to enable a view of proteins working together in the same pathway. These statistics include pathway coverage, odds ratio and the corresponding *P*-value calculated by Fisher's exact test (12).

Versatile query and search tools enable complex queries to explore these experiment-level pathway datasets. Under the Pathways tab, researchers can query for pathways by full or partial names as well as by bimolecular identifiers. The pathway expression results can then be sorted or downloaded. For example, the user can visualize the results for pathways expressed in cystic fibrosis-related experiments (Figure 3).

Experimental metadata

Under the Experiments tab, researchers can search, filter and review experimental metadata. For example, querying for 'cystic fibrosis' will return a list of experiments where the search term appears in the experiment description. The results include summary statistics on genes, proteins and pathways expressed in each experiment.

Experimental metadata are captured using a multi-omics checklist (9–11,45). This checklist provides information about experimental design, data collection and analysis. This added layer of experimental metadata can be used by researchers to assess the quality, consistency, usability and reproducibility of data in MOPED.

MOPED currently contains experimental metadata checklists on 12 proteomics and all 183 transcriptomics experiments (9,10). For example, the *snyder_personal_omics_profiling* experiment has experimental metadata describing experiment design, data collection and data analysis (Supplementary Figure S2) (45). Built-in links to GEO provide additional information about the transcriptomics experiments (21). The MOPED team is currently working on completing metadata checklists for all experiments in MOPED.

Visualizations

MOPED provides interactive visualizations including bar graphs of protein, gene and pathway expression levels across experiments that can be viewed by either tissue or condition (Figures 1 and 3). The expression matrix compares expression of neighboring genes and corresponding proteins along the chromosome that can be grouped by tissue or condition. The chord diagram gives an overview of MOPED data content for different organisms. Mouse-over descriptions enhance the plots by providing more detailed graphical and experimental information.

MOPED'S IMPACT

MOPED provides support for the scientific community as a resource of processed expression data. Researchers have used MOPED to help identify potential proteins to study based on expression levels (46), to compare expression of different proteins across experiments, conditions, localizations and tissues (47,48), to investigate expression levels in various tissue types, cell lines, conditions and diseases (49–51) and to help link uncharacterized proteins to diabetes and cancer (52,53).

MOPED was originally created in response to a survey indicating a need for a processed expression database (1,3). Through direct and indirect interactions, the MOPED team is able to learn the needs of biological and biomedical researchers and respond with updates and improvements. With this feedback in mind, MOPED acknowledges the changing needs of the life sciences community.

Data submissions

To be a complete and accurate data resource, MOPED needs to reflect the current understanding of gene, protein and pathway expression patterns through access to the most recent data and experiments. MOPED provides a convenient venue for researchers to both make their data public and share these privately with collaborators and reviewers (1,3). Private MOPED provides the same features, tools and conveniences as the public MOPED but enables the owner of the data to both limit access and make the data public when appropriate. If interested in submitting data to MOPED, please contact the MOPED team through the Forum: <http://moped-forum.proteinspire.org>.

Future directions

As the field of biology rapidly advances, MOPED is advancing with it. MOPED took the following steps to complement developments in research: transitioned from proteomics to multi-omics (8), added metadata to enable more reliable use and integration of multiple data sources (9,11) and used pathways to create a coherent view of multi-omics data and an experimental level assessment of expression (12). The next steps for MOPED are to enhance integration of different MOPED data types. This will be done by providing integrated data views, graphical displays and pathway analyses that show and utilize multiple omics data types. In addition, future versions of MOPED

Table 3. MOPED references the following external resources on the Protein and Gene Details pages (13–16,33–43).

External source	Protein	Gene	Reference
GeneCards*	X	X	(33)
PDB*	X		(41,42)
BioCyc*	X		(16)
UniProt	X		(13)
NCBI Gene	X	X	(15)
RefSeq	X		(34)
ensembl	X	X	(14)
EMBL	X		(35)
WormBase	X	X	(36)
MGI	X		(37)
KEGG	X		(38)
HGNC	X	X	(39)
SGD	X		(40)
GenBank		X	(43)
GO	X	X	(44)

* indicates resources that are cross-referenced with MOPED.

will offer disease-centric data integration and include additional model organisms. The disease-centric view will allow the comparison of biomolecular processes and their functions between normal and disease states. Also, MOPED plans to add pathway relative expression statistics to identify over/under-expression of pathways in relative expression experiments for both proteomics and transcriptomics data (54,55). These steps will continue the advancement of MOPED as a resource for biological and biomedical science.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENT

We thank Maggie Lackey for her critical reading the manuscript.

FUNDING

National Science Foundation under the Division of Biological Infrastructure [0969929]; The Robert B. McMillen Foundation; Seattle Children’s Research Institute [to E.K.]. Funding for open access charge: Seattle Children’s Research Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Higdon,R., Stewart,E., Stanberry,L., Haynes,W., Choiniere,J., Montague,E., Anderson,N., Yandl,G., Janko,I., Broomall,W. et al. (2014) MOPED enables discoveries through consistently processed proteomics data. *J. Proteome Res.*, **13**, 107–113.
2. Kolker,E., Higdon,R., Haynes,W., Welch,D., Broomall,W., Lancet,D., Stanberry,L. and Kolker,N. (2012) MOPED: Model Organism Protein Expression Database. *Nucleic Acids Res.*, **40**, D1093–D1099.
3. Higdon,R., Haynes,W., Stanberry,L., Stewart,E., Yandl,G., Howard,C., Broomall,W., Kolker,N. and Kolker,E. (2013) Unraveling the complexities of life sciences data. *Big Data*, **1**, 42–50.

4. Zhang,W., Li,F. and Nie,L. (2010) Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiol. Read. Engl.*, **156**, 287–301.
5. Sikaroodi,M., Galachiantz,Y. and Baranova,A. (2010) Tumor markers: the potential of ‘omics’ approach. *Curr. Mol. Med.*, **10**, 249–257.
6. Stanberry,L., Mias,G.I., Haynes,W., Higdon,R., Snyder,M. and Kolker,E. (2013) Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. *Metabolites*, **3**, 741–760.
7. Chen,R., Mias,G.I., Li-Pook-Than,J., Jiang,L., Lam,H.Y.K., Chen,R., Miriami,E., Karczewski,K.J., Hariharan,M., Dewey,F.E. et al. (2012) Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell*, **148**, 1293–1307.
8. Montague,E., Stanberry,L., Higdon,R., Janko,I., Lee,E., Anderson,N., Choiniere,J., Stewart,E., Yandl,G., Broomall,W. et al. (2014) MOPED 2.5—an integrated multi-omics resource: Multi-Omics Profiling Expression Database now includes transcriptomics data. *Omics J. Integr. Biol.*, **18**, 335–343.
9. Kolker,E., Özdemir,V., Martens,L., Hancock,W., Anderson,G., Anderson,N., Aynacioglu,S., Baranova,A., Campagna,S.R., Chen,R. et al. (2014) Toward more transparent and reproducible omics studies through a common metadata checklist and data publications. *Omics J. Integr. Biol.*, **18**, 10–14.
10. Dumbill,E. and Kolker,E. (2013) Introducing a metadata checklist for omics data. *Big Data*, **1**, 195–195.
11. Announcement: Reducing our irreproducibility (2013) *Nature*, **496**, 398–398.
12. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
13. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
14. Flicek,P., Amodé,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. et al. (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
15. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
16. Caspi,R., Altman,T., Billington,R., Dreher,K., Foerster,H., Fulcher,C.A., Holland,T.A., Keseler,I.M., Kothari,A., Kubo,A. et al. (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Res.*, **42**, D459–D471.
17. Mi,H., Muruganujan,A. and Thomas,P.D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.*, **41**, D377–D386.
18. Croft,D., Mundo,A.F., Haw,R., Milacic,M., Weiser,J., Wu,G., Caudy,M., Garapati,P., Gillespie,M., Kamdar,M.R. et al. (2014) The

- Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, D472–D477.
19. Vizcaino, J.A., Côté, R.G., Csordas, A., Dianes, J.A., Fabregat, A., Foster, J.M., Griss, J., Alpi, E., Birim, M., Contell, J. *et al.* (2013) The PRoteomics IDentifications (PRIDE) database and associated tools: status in 2013. *Nucleic Acids Res.*, **41**, D1063–D1069.
 20. Desiere, F. (2006) The PeptideAtlas project. *Nucleic Acids Res.*, **34**, D655–D658.
 21. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
 22. Kolker, E., Higdón, R., Welch, D., Bauman, A., Stewart, E., Haynes, W., Broomall, W. and Kolker, N. (2011) SPIRE: systematic protein investigative research environment. *J. Proteomics*, **75**, 122–126.
 23. Higdón, Roger, van Belle, Gerald and Kolker, Eugene (2008) A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics (Oxford, England)*, **24**, 1225–1228. [PubMed]
 24. Hather, G., Higdón, R., Bauman, A., von Haller, P.D. and Kolker, E. (2010) Estimating false discovery rates for peptide and protein identification using randomized databases. *Proteomics*, **10**, 2369–2376.
 25. Higdón, R., Reiter, L., Hather, G., Haynes, W., Kolker, N., Stewart, E., Bauman, A.T., Picotti, P., Schmidt, A., van Belle, G. *et al.* (2011) IPM: an integrated protein model for false discovery rate estimation and identification in high-throughput proteomics. *J. Proteomics*, **75**, 116–121.
 26. Holzman, T. and Kolker, E. (2004) Statistical analysis of global gene expression data: some practical considerations. *Curr. Opin. Biotechnol.*, **15**, 52–57.
 27. Smyth, G.K. (2005) Limma: linear models for microarray data. In: Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. and Huber, W. (eds). *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, NY, pp. 397–420.
 28. Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S. *et al.* (2014) A draft map of the human proteome. *Nature*, **509**, 575–581.
 29. Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H. *et al.* (2014) Mass-spectrometry-based draft of the human proteome. *Nature*, **509**, 582–587.
 30. Geiger, T., Madden, S.F., Gallagher, W.M., Cox, J. and Mann, M. (2012) Proteomic portrait of human breast cancer progression identifies novel prognostic markers. *Cancer Res.*, **72**, 2428–2439.
 31. Moghaddas Gholami, A., Hahne, H., Wu, Z., Auer, F.J., Meng, C., Wilhelm, M. and Kuster, B. (2013) Global proteome analysis of the NCI-60 cell line panel. *Cell Rep.*, **4**, 609–620.
 32. Pfister, T.D., Reinhold, W.C., Agama, K., Gupta, S., Khin, S.A., Kinders, R.J., Parchment, R.E., Tomaszewski, J.E., Doroshow, J.H. and Pommier, Y. (2009) Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. *Mol. Cancer Ther.*, **8**, 1878–1884.
 33. Stelzer, G., Dalah, I., Stein, T.I., Satanower, Y., Rosen, N., Nativ, N., Oz-Levi, D., Olender, T., Belinky, F., Bahir, I. *et al.* (2011) In-silico human genomics with GeneCards. *Hum. Genomics*, **5**, 709–717.
 34. Pruitt, K.D., Brown, G.R., Hiatt, S.M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C.M., Hart, J., Landrum, M.J., McGarvey, K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
 35. Pakseresht, N., Alako, B., Amid, C., Cerdeño-Tárraga, A., Cleland, I., Gibson, R., Goodgame, N., Gur, T., Jang, M., Kay, S. *et al.* (2014) Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res.*, **42**, D38–D43.
 36. Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K. *et al.* (2014) WormBase 2014: new views of curated biology. *Nucleic Acids Res.*, **42**, D789–D793.
 37. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E. and Mouse Genome Database Group. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, **42**, D810–D817.
 38. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
 39. Gray, K.A., Daugherty, L.C., Gordon, S.M., Seal, R.L., Wright, M.W. and Bruford, E.A. (2013) Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545–D552.
 40. Costanzo, M.C., Engel, S.R., Wong, E.D., Lloyd, P., Karra, K., Chan, E.T., Weng, S., Paskov, K.M., Roe, G.R., Binkley, G. *et al.* (2014) *Saccharomyces* genome database provides new regulation data. *Nucleic Acids Res.*, **42**, D717–D725.
 41. Rose, P.W., Bi, C., Bluhm, W.F., Christie, C.H., Dimitropoulos, D., Dutta, S., Green, R.K., Goodsell, D.S., Prlic, A., Quesada, M. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
 42. Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Fernandez Montecelo, M.A., van Ginkel, G., Gore, S.P. *et al.* (2014) PDBE: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.
 43. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2013) GenBank. *Nucleic Acids Res.*, **41**, D36–D42.
 44. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
 45. Snyder, M., Mias, G., Stanberry, L. and Kolker, E. (2014) Metadata checklist for the integrated personal OMICS study: proteomics and metabolomics experiments. *Omics J. Integr. Biol.*, **18**, 81–85.
 46. Xue, J., Wijeratne, S.S.K. and Zemleni, J. (2013) Holocarboxylase synthetase synergizes with methyl CpG binding protein 2 and DNA methyltransferase 1 in the transcriptional repression of long-terminal repeats. *Epigenetics*, **8**, 504–511.
 47. Staneva, R., Rukova, B., Hadjidekova, S., Nesheva, D., Antonova, O., Dimitrov, P., Simeonov, V., Stamenov, G., Cukuranovic, R., Cukuranovic, J. *et al.* (2013) Whole genome methylation array analysis reveals new aspects in Balkan endemic nephropathy etiology. *BMC Nephrol.*, **14**, 225.
 48. Wilhelm, T. (2014) Phenotype prediction based on genome-wide DNA methylation data. *BMC Bioinformatics*, **15**, 193.
 49. Blacker, T.S., Mann, Z.F., Gale, J.E., Ziegler, M., Bain, A.J., Szabadkai, G. and Duchon, M.R. (2014) Separating NADH and NADPH fluorescence in live cells and tissues using FLIM. *Nat. Commun.*, **5**, 3936.
 50. Williams, B.C., Filter, J.J., Blake-Hodek, K.A., Wadzinski, B.E., Fuda, N.J., Shalloway, D. and Goldberg, M.L. (2014) Greatwall-phosphorylated Endosulfine is both an inhibitor and a substrate of PP2A-B55 heterotrimer. *eLife*, **3**, e01695.
 51. Chen, M. and Penning, T.M. (2014) 5 β -reduced steroids and human $\Delta(4)$ -3-ketosteroid 5 β -reductase (AKR1D1). *Steroids*, **83**, 17–26.
 52. Delgado, A.P., Brandao, P. and Narayanan, R. (2014) Diabetes associated genes from the dark matter of the human proteome. *MOJ Proteomics Bioinform.*, **1**, 00020.
 53. Delgado, A.P., Brandao, P., Chapado, M.J., Hamid, S. and Narayanan, R. (2014) Open reading frames associated with cancer in the dark matter of the human genome. *Cancer Genomics Proteomics*, **11**, 201–213.
 54. Haynes, W.A., Higdón, R., Stanberry, L., Collins, D. and Kolker, E. (2013) Differential expression analysis for pathways. *PLoS Comput. Biol.*, **9**, e1002967.
 55. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 15545–15550.